SGN-2500 Johdatus hahmontunnistukseen
Tentti 1 18.3.2009

Huomaa, että laskutehtävissä oikeat vastaukset ilman selitystä kuinka ne ollaan saatu tuottavat nolla pistettä, joten selitä miten laskit vastaukset! Muista myös määritellä KAIKKI vastauksissa esiintyvät symbolit. Laskimen käyttö on sallittua.

1. Selitä seuraavat kolme ohjattujen luokitinten virhelähdettä :

   (a) Bayes virhe

   (b) Mallivirhe

   (c) Estimointivirhe

   Miten ML-estimaatteihin perustuvassa luokituksessa näitä virheitä voitaisiin pienentää? (6p)

2. Ajatellaan kahden kategorian ja yhden piirteen luokitusongelmaa. Oletetaan, että kategoria ω_1 on eksponentiaalisesti jakautunut ja kategoria ω_2 on normaalisti jakautunut. On olemassa sekoiteotannalla kerätty opetusdata

   $$\mathcal{D}_1 = \{4.78, 4.84, 1.51, 3.90, 4.84\}, \mathcal{D}_2 = \{5.19, 5.18, 3.96, 4.32, 4.17, 4.50\}.$$

   Opeta Bayes luokitin perustuen suurimman uskottavuuden estimaatteihin ja luokita piirteet $x = 4$ ja $x = 8$. Muista myös tarkastella prioritodennäköisyyksiä!(6p)
   Eksponentiaalijakauman tiheysfunktio on

   $$p(x|\theta) = \begin{cases} \theta \exp(-\theta x) & \text{jos} \quad x \geq 0 \\ 0 & \text{muulloin} \end{cases}$$

   ja suurimman uskottavuuden (maximum-likelihood, ML) estimaatti parametrille θ on

   $$\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}.$$

   Normaalijakauman tiheysfunktio on $p_{normal}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2}(\frac{(x-\mu)^2}{\sigma^2})]$. Parametrin μ ML-estimaatti on otoskeskiarvo ja parametrin σ^2 ML-estimaatti on otosvarianssi.

3. Ajatellaan kolmen kategorian k:n lähimmän naapurin luokittelijaa ja seuraavia neljän piirteen harjoitusnäytteitä

ω_1	(-1,0,0,-1)	(1, 1,-2,0)	(1,3,1,1)	(0,-1,-1,2)	(0,0,0,0)
ω_2	(2,0,2,2)	(1,4,1,1)	(2,2,2,1)	(1,2,-1,0)	(2,0,1,2)
ω_3	(3,2,4,2)	(3,1,3,2)	(1,4,2,6)	(3,4,4,5)	(2,3,3,3)

   Luokita piste (1,2,1,1) perustuen 3-lähimmän naapurin sääntöön
   (a) käyttäen etäisyysmittana Euklidista etäisyyttä (3p)
   (b) käyttäen etäisyysmittana L_1 etäisyyttä (3p). **Vinkki**: L_1 etäisyys pisteiden **a** ja **b** välillä on $L_1(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{d} |a_i - b_i|$.

4. Ajatellaan kahden luokan lineaarista luokitinta, jonka erotinfunktio on $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$. Osoita, että päätöspinnan normaali on painovektori **w**:n suuntainen. (6p)

5. (a) Selitä ohjatun ja ohjaamattoman oppimisen ero (2p).
   (b) Tarkastellaan ohjaamatonta oppimista k-means kriteerin avulla. Olkoon $\mathbf{x}_1 = (3,5)^T, \mathbf{x}_2 = (2,4)^T, \mathbf{x}_3 = (1,0)^T, \mathbf{x}_4 = (5,0)^T$, ja ajatellaan seuraavaa kolmea partitiota:

   (a) $\mathcal{D}_1 = \{\mathbf{x}_1, \mathbf{x}_2\}, \mathcal{D}_2 = \{\mathbf{x}_3, \mathbf{x}_4\}$

   (b) $\mathcal{D}_1 = \{\mathbf{x}_1, \mathbf{x}_4\}, \mathcal{D}_2 = \{\mathbf{x}_2, \mathbf{x}_3\}$

   (c) $\mathcal{D}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}, \mathcal{D}_2 = \{\mathbf{x}_4\}$

   Laske k-means-kriteerin (eli neliösummakriteerin) arvot näille partitioille. Mitä partitiota k-means kriteeri suosii? (4p)